HAND DETECTION ON IMAGES BASED ON DEFORMABLE PART MODELS AND ADDITIONAL FEATURES

BY

PEDRO MOISES CRISOSTOMO ROMERO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Adviser:

Professor David Forsyth

Abstract

Hand detection on images has important applications on person activities recognition. This thesis focuses on PASCAL Visual Object Classes (VOC) system for hand detection. VOC has become a popular system for object detection, based on twenty common objects, and has been released with a successful deformable parts model in VOC2007. A hand detection on an image is made when the system gets a bounding box which overlaps with at least 50% of any ground truth bounding box for a hand on the image. The initial average precision of this detector is around 0.215 compared with a state-of-art of 0.104; however, color and frequency features for detected bounding boxes contain important information for re-scoring, and the average precision can be improved to 0.218 with these features. Results show that these features help on getting higher precision for low recall, even though the average precision is similar.

To my parents Sabina and Juan

Acknowledgments

Thanks to the Fulbright Commission and the University of Illinois at Urbana-Champaign for supporting my studies and indeed this work.

Table of Contents

Chapter 1

Introduction

Hand detection on images has many potential applications, such as Human Computer Interaction (robotics, smartphones control, games consoles), Sign Language Recognition (deaf and dumb language, teamwork communication, instructions), and Human Action Recognition (object manipulation). Usually the first two applications require real time systems where the background is not complex and a hand is tracked. On the other hand, for the third application the image should contain a person doing something with a variety of backgrounds and without hand tracking. To detect a hand in these conditions is a bigger challenge.

A hand is a very complex object because it could adopt multiple poses and projections on images. Some ideas to solve this problem include 3D hand pose estimation [1], hand tracking using temporal information [5, 6], or color analysis [8]. These methods have assumptions such as presence of hand, moderated clean background, temporal sequence of hand images, partial or no occlusion, or lightning conditions.

We propose a general hand detector on still images which could have multiple hands (even partially occluded) or none. Also, these images could have any background and light conditions. The only requirement is a minimum resolution for hands.

The object detection goal is to find an object location and size of a pre-defined class in an image or video in spite of wide variations in visual appearance due to changes in the form and color of the object, occlusions, geometrical transformations (such as scaling and rotation), changes in illumination, and potentially non-rigid deformations of the object itself [7]. In recent years many studies, both in machine learning and computer vision areas, have focused on this problem. Since detailed hand-segmentation and labelling of images is very labour intensive, learning object categories from weakly labeled data has been studied in recent years. Weakly labeled data means that training images are labelled only according to the presence or absence of each category of object. A major challenge presented by this problem is that the foreground object is accompanied by widely varying background clutter, and the system must learn to distinguish the

foreground from the background without the aid of labeled data. Many approaches to object recognition are founded on probability theory, and can be broadly characterized as either generative or discriminative according to whether or not the distribution of the image features is modeled. Generative and discriminative methods have very different characteristics, as well as complementary strengths and weaknesses.

An object detection method using part based models proposed by Felzenswalb et al. [3] has became popular after showing a good performance on The PASCAL Visual Object Classes (VOC) Challenge. Many researchers are using this system as a black box for new object detection algorithms. Even when hand is not one of the twenty classes of object in PASCAL database, we think this system could also work for hands.

Chapter 2

Related Work

As we mentioned in Chapter 1, most hand detection systems work with some restrictions. Because we want to detect hands in general images, we use the part based models proposed by Felzenswalb et al. [3] as a main component for our hand detector and other detectors based on color and frequency features. Also, we follow the directions given by the PASCAL VOC Challenge with respect to annotations and functions to compare our results with those reported on this competition.

2.1 PASCAL Visual Object Classes (VOC) Challenge

The PASCAL Visual Object Classes (VOC) Challenge [2] is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard database of images and annotation, and standard evaluation procedures. Organized annually from 2005 to present, the challenge and its associated dataset has become accepted as the benchmark for object detection.

The VOC 2010 Challenge goal is to recognize objects from a number of visual object classes in realistic scenes (i.e. not pre-segmented objects). Fundamentally, it is a supervised learning problem where a training set consists on a set of labeled images. The twenty object classes that have been selected are:

- Person: person
- Animal: bird, cat, cow, dog, horse, sheep
- Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
- Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

There are three main competitions: classification, detection, and segmentation; and three "taster" competition: person layout, action classification, and ImageNet large scale recognition.

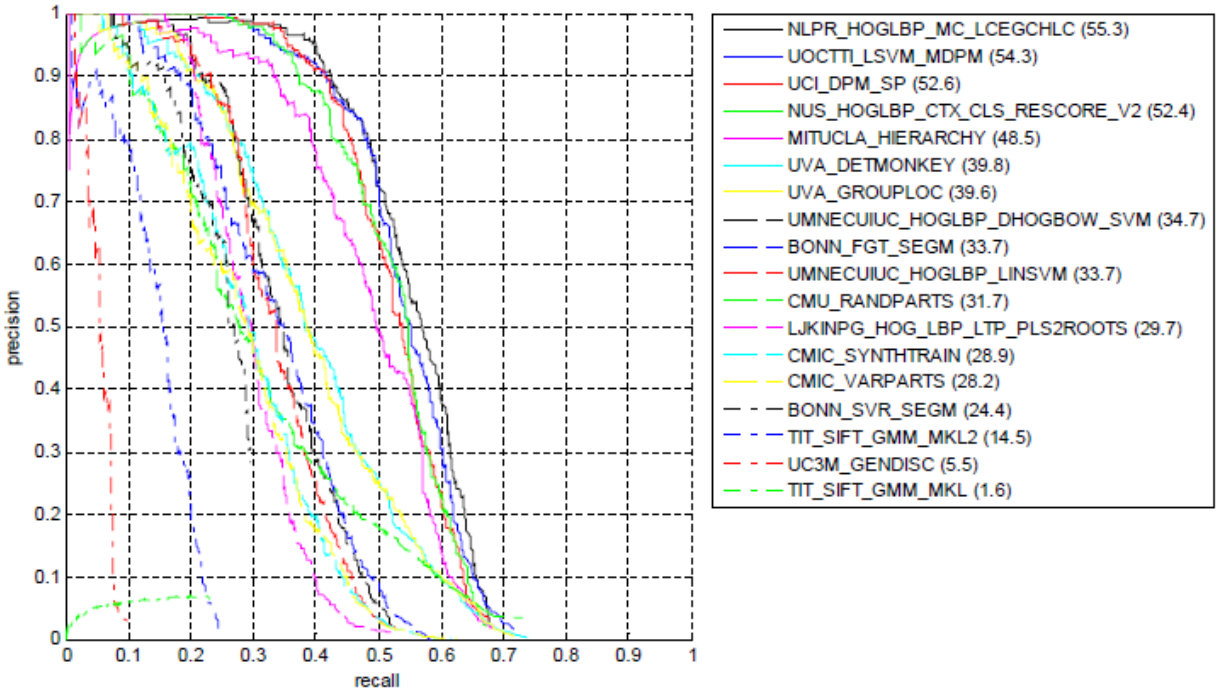

Figure 2.1: Precision - recall curve for bycicle detection on the PASCAL VOC2010 challenge. Taken from PASCAL VOC2010 results webpage.

Figure 2.1 shows precision-recall curves for bycicle detection for all competitors on VOC2010. The winner raise an average precision of 0.553. Currently, the best average precision is 0.584 for aeroplane and the worse one is 0.130 for potted plant.

The Person layout challenge consists on predicting bounding boxes for a person and its parts (head, hands, feet). Figure 2.2 shows the results for hand detectors on the VOC2010 Person Layout competition using own data from competitors. The winner is the Oxford Skin Based Detector (Oxford_SBD) with an average precision of 0.104 and the runner up (team BCNPCL_HumanLayout from Barcelona) got an average precision of 0.033. The algorithm for the Oxford_SBD method basically consists on a skin detector and analysis of isolated blobs or blobs at the end of arms.

Felzenszwalb et al. proposed a complete learning-based system for detecting and localizing objects in images, which is a core component for most current best detectors [3]. The system represents objects using mixtures of deformable part models. These models are trained using a discriminative method that only requires bounding boxes for the objects in an image. The approach leads to efficient object detectors that

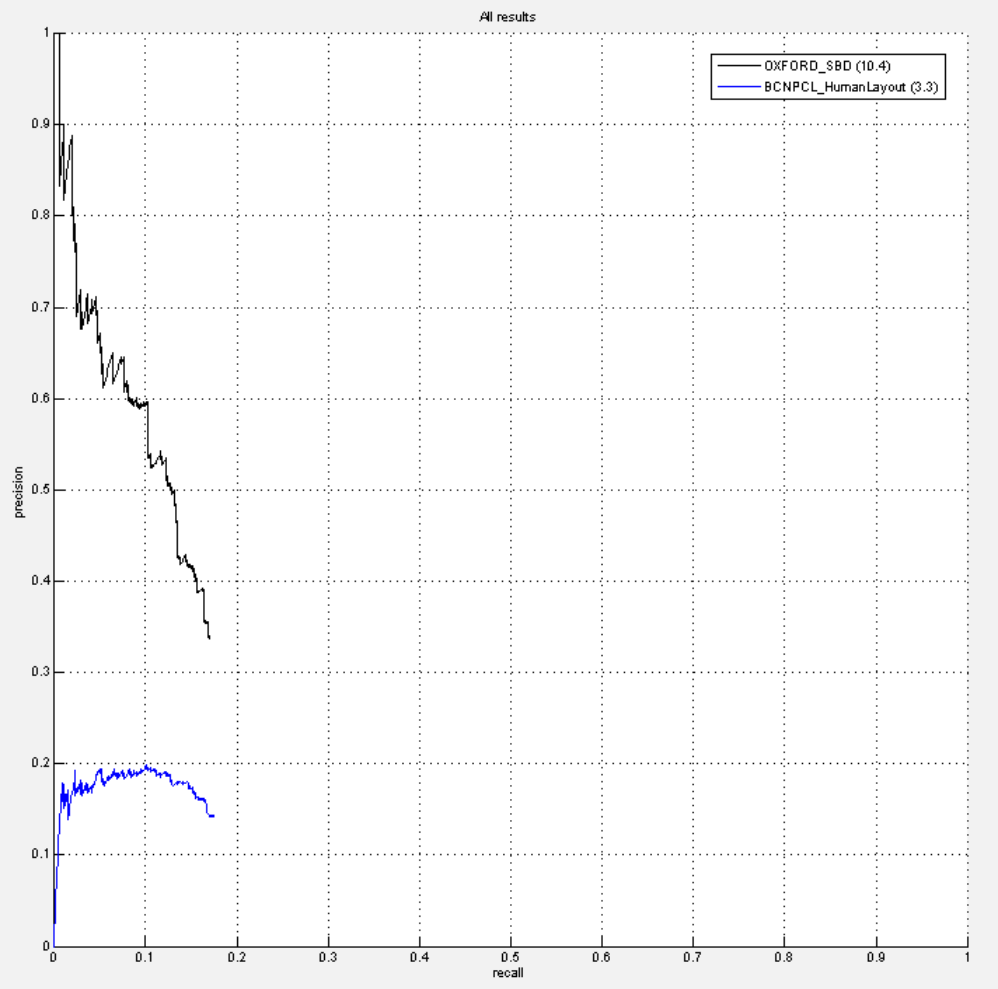

Figure 2.2: Precision - recall curve for the PASCAL VOC2010 Person Layout (class "hand") competitors. The average precision for the winner (Oxford) was 0.104 while the runner up's (Barcelona team) was 0.033. Taken from PASCAL VOC2010 results webpage.

achieve state-of-the-art results on the PASCAL and INRIA person datasets.

At a high level, this system can be characterized by the combination of:

- Strong low-level features based on histograms of oriented gradients (HOG).
- Efficient matching algorithms for deformable part-based models (pictorial structures).
- Discriminative learning with latent variables (latent SVM).

Chapter 3

Hand Detection Description

We start with a detector based on a deformable parts model that gives us a set of bounding boxes with a confidence value for each of them. In order to improve this detector, we propose the use of color and frequency features. We test the following 4 detectors:

- Detector 1: Baseline detector based on VOC 2007 toolkit, which implements a deformable parts model for learning objects.
- Detector 2: Includes Detector 1 and a detector using **color** features.
- Detector 3: Includes Detector 1 and a detector using **frequency** features.
- Detector 4: Includes Detector 1 and a detector using both **color** and **frequency** features.

Some considerations for images:

- Color images in JPEG format with 8 bits per planes red, green and blue (RGB).
- Image resolution in order to have hands with at least 60 pixels per side for positive examples.
- Hands in any direction, free or holding objects, including partially occluded fingers.

3.1 Detector 1: VOC2007 Development Kit

This package contains the implementation of Discriminatively Trained Deformable Part Models. VOC2007 dataset and annotations are available to test this code for 20 objects, not including hands. Because the hand resolution in this dataset is low, a new dataset of higher resolution images (containing at least one hand) were collected from Flickr and added to the VOC2007 dataset. We have chosen images where a bounding box for a hand has a minimum of 60 pixels per side. The training set has 250 positive and 2508 negative

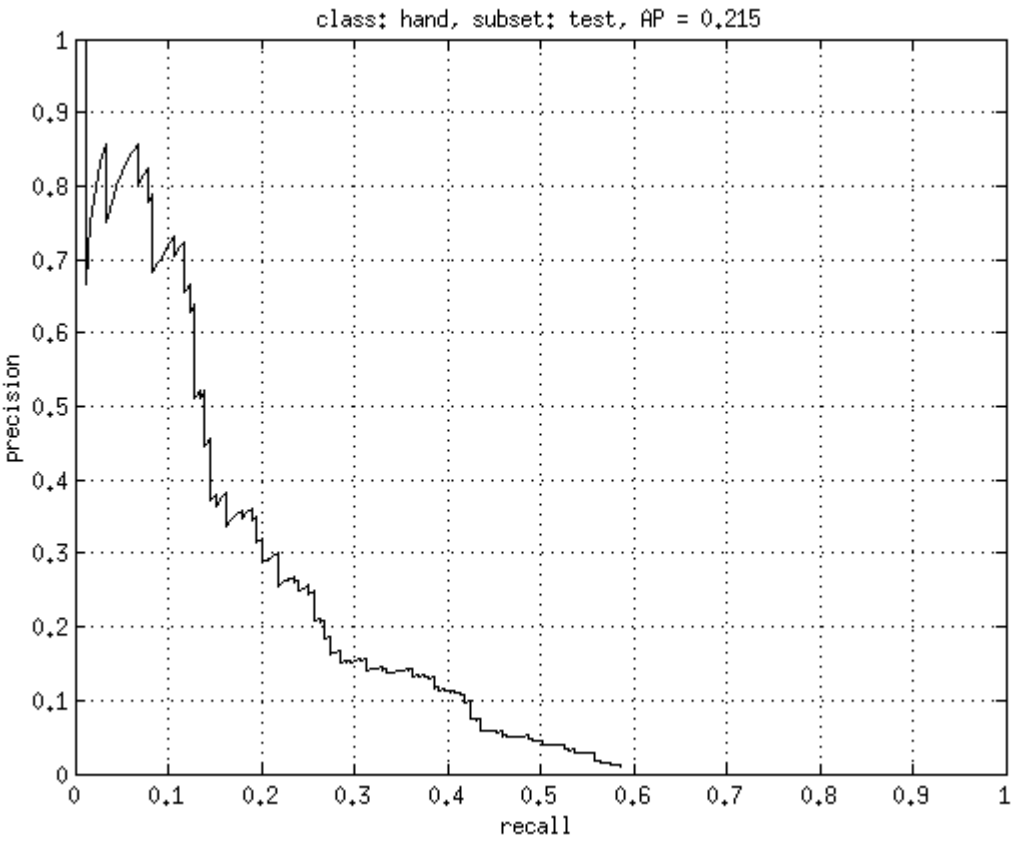

Figure 3.1: Precision - recall curve for test data using Detector 1. Our average precision is 0.215 while the PASCAL VOC2010 winner's is 0.104.

examples and the testing set has 100 positive and 602 negative examples. We also added annotations for these positive examples of hands, while any image from VOC2007 dataset (which could contain bounding boxes for any of the twenty original classes) is considered negative example.

A training and evaluating process for a hand model is performed with this new dataset in the manner recommended by VOC2007. The best average precision obtained was 0.215 for a hand model with 6 components. Figure 3.1 shows the precision-recall graph. The best average precision reported by the VOC2007 detections challenge for all twenty objects is in the range from 0.10 to 0.43, whereas for VOC2010 these values are in the range from 0.13 to 0.60. Because a hand is more deformable than all those twenty objects an average precision of 0.215 is acceptable. Figures 3.2(a) and 3.2(b) show examples of bounding boxes for detected hands and false positives, both with high confidence value.

In order to improve the previous result, we propose a new model using the hand detector confidence value and new features from bounding boxes. We use the following features:

(a) Detected hands with high confidence.

(b) False positive hands with high confidence.

Figure 3.2: Bounding boxes examples using Detector 1 (baseline)

- Color analysis, because many false positive bounding boxes contain sections with colors far from skin color.

- Frequency analysis, because high contrast on finger boundaries contributes to a specific range of spatial

frequency even when fingers are oriented in different directions or partially occluded.

One alternative that proved unhelpful was a body parts (arm, elbow, leg, knee) detector. Because some false positives include body parts and they are expected to appear close to hands, we tried to detect them in image sections a little bit bigger than the detected bounding box. This was implemented using the same development kit, however this idea was discarded because these body parts are too simple and they do not hold enough information to be learn. In general, for most bounding boxes, the system detected body parts anywhere; for instance, fingers were detected as arms.

3.2 Detector 2: Color Analysis

A bounding box containing a hand is expected to have many pixels with intensities in every color plane related to skin color. Robust histograms based on rg and by variables were used as color features. The process follows these steps:

- Take only the 25% of pixels which correspond to the central part of the bounding box. Figure 3.3 shows two bounding boxes for a detected hand and a false positive, with the region of interest for color analysis.

- Calculate a robust color histogram using two new variables **rg** and **by** which correspond to subtraction of color planes (r-g and b-y), where r, g, b and y represents color planes red, green, blue and the mean of r and g planes.

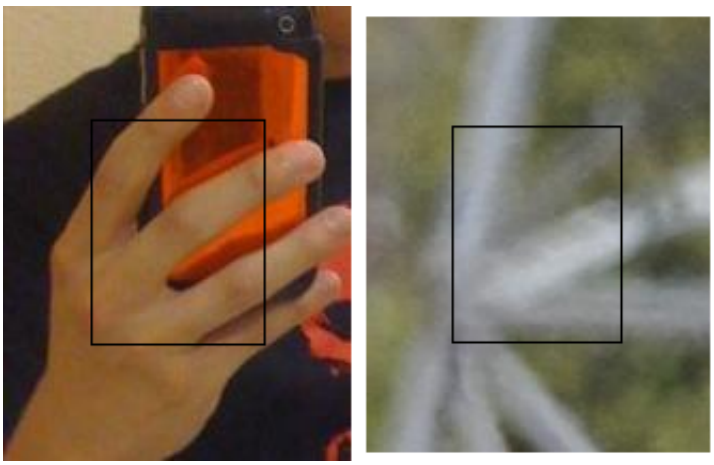

Figure 3.3: Both images are bounding boxes with hand (left) and false positive (right) from Detector 1 where only the central region limited by black lines are used for color analysis.

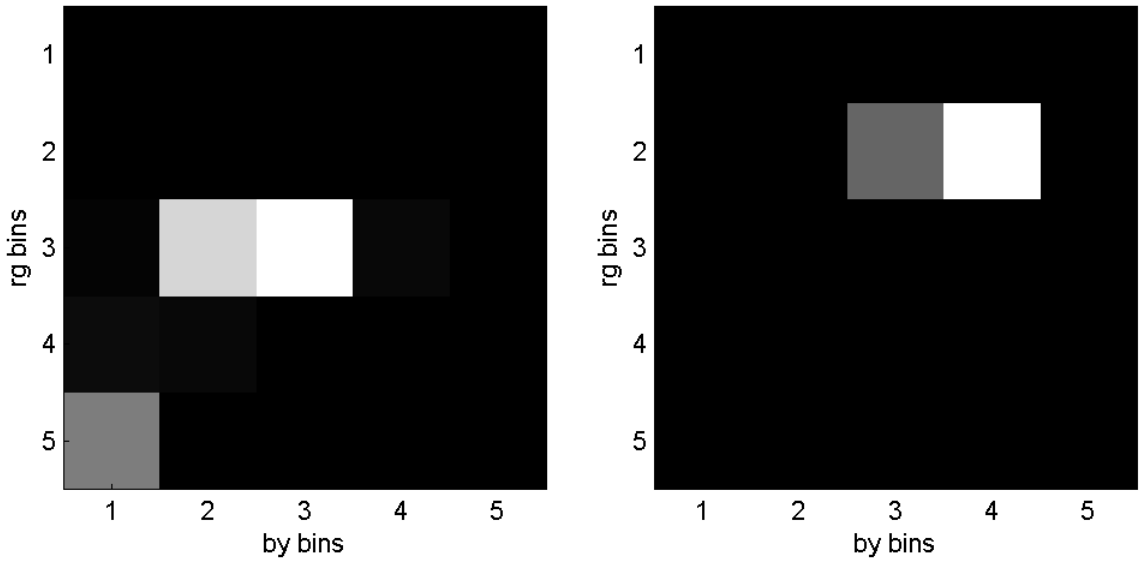

Figure 3.4: Two dimensional histograms for the true (left) and false (right) positive bounding boxes shown in Figure 3.3. Brighter gray means high proportion of bounding boxes pixels in the respective bin.

- Using 100 bounding boxes with hands from training data we calculate the mean μ and standard deviation σ for each variable and all pixels from those images. These values were $\mu_{rg} = 0.1332, \sigma_{rg} = 0.1186, \mu_{by} = -0.1174, \sigma_{by} = 0.0994$. Then, for a bounding box with a hand we expect to have many pixels close to the mean for both variables; i.e. higher value at the center of its color histogram.

- Only five bins are considered for each variable which give us a total of 25 bins for this two dimensional histogram. Bins are limited by $\mu - 2\sigma, \mu - \sigma, \mu + \sigma, \mu + 2\sigma$ values. Figure 3.4 shows 2D histograms for images on Figure 3.3 where gray intensity represents the relative amount of pixels for each bin; brighter gray means high proportion of bounding boxes pixels in the respective bin. Notice that for the first histogram (left), which corresponds to a hand, the higher value is in the center as we expected, whereas for the second histogram (right) the higher value is not centered because the mean color is not close to the skin color.

- Get a model of color histograms using a Kernel SVM [4] for some labeled bounding boxes from training data. Figure 3.5 shows scores for positive and negative bounding boxes from training data. We can see that bounding boxes with a low color score (around -1) and medium confidence (less than -0.55) can be dropped.

- Get a score for all bounding boxes from testing data using the model found in previous step and drop all bounding boxes with color score below -0.95.

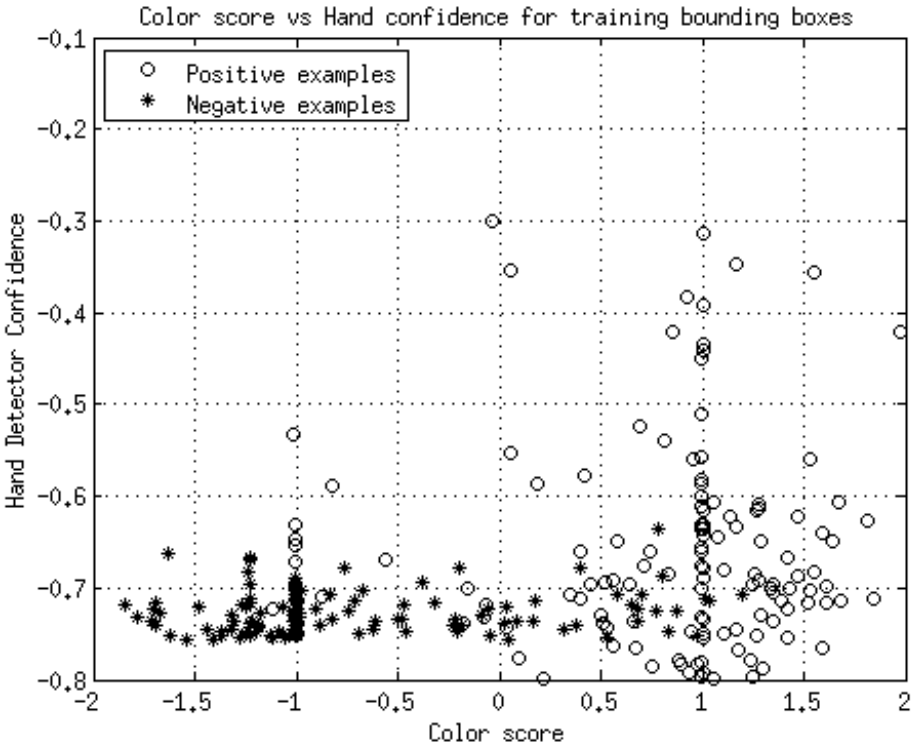

Figure 3.5: Positive and negative examples from training data

.

The precision-recall curve after applying the color features to initial bounding boxes from testing data is shown in Figure 3.6. Notice that the average precision (0.202) is lower than the original (0.215), however we have higher precision for low recall. Also, we can see this detector have lower precision for recall value from 0.10 to 0.13.

3.3 Detector 3: Frequency Analysis

A bounding box containing a hand is expected to show fingers, which have a constant width and relative high contrast at their edges. This means that fingers hold information for a specific spatial frequency. We notice visually that grayscale images, obtained from color images, still have high contrast at finger boundaries; for this reason we transform color images to grayscale in order to perform this frequency analysis. Because the orientation and real size of fingers are unknown, we propose the use of multiscale directional filters to find the energy related to fingers.

Robust histograms based on total energy per scale were used as a frequency feature. The process follows

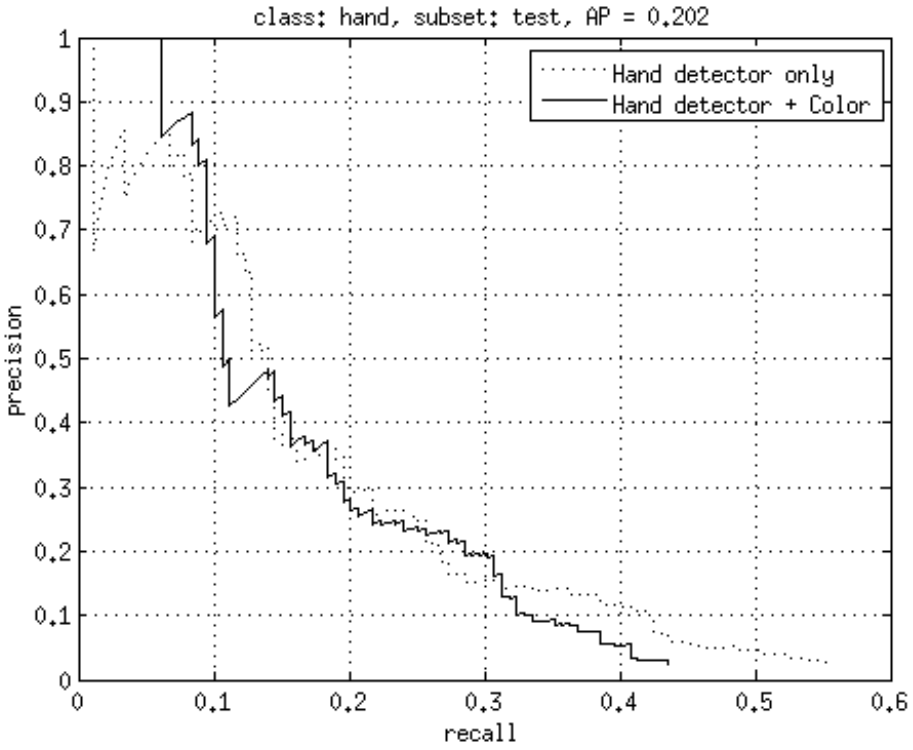

Figure 3.6: Precision - recall curve on test data including color analysis. The average precision for this detector is 0.202 while the average precision for detector 1 is 0.215.

.

these steps:

- Build a set of 12 directional filters (every 15 degrees). Since we expect to have, for the minimum dimension of the bounding box, from 4 to 12 fingers, we work with 5 scales. Figure 3.7 shows a set of 12 filters for one scale. Notice that they are not symmetric because we also actually used a mirrored filter to complement the response and ensure to find regions with edges at both sides that correspond to fingers edges. These filters should respond to dark contrast on finger boundaries. We found these shape experimentally since we were focused on detect the high contrast at finger edges, even when we have fingers touching each other and showing narrow edges. The filters are circular with positive values for all positions inside the circle except on a segment with negative values, The summation of all positives values is one and for all negatives is -1; outside the circle the value is zero. It is better when we have separated fingers because it usually guarantees more pixels with high contrast for fingers intensity values. We repeat this process for every scale where the filter size for next scale is calculated as the division of previous one by fifth square of 3.

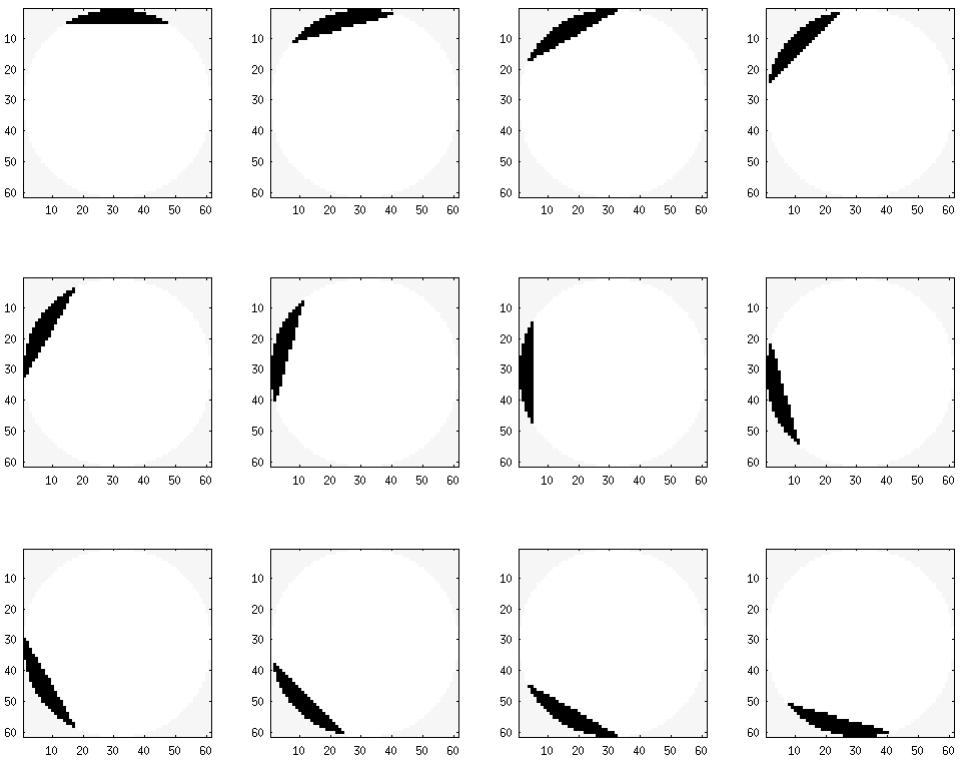

Figure 3.7: Directional filters for one scale at every 15 degrees starting at 0 degrees with respect to the horizontal axis.

- For every scale get one image as follow: get 12 images after applying all directional filters to the grayscale bounding box; for every position on the image get the maximum value from all 12 filter responses. This gives us the maximum energy at every pixel for any direction. Actually, we consider only no negative energy values. Figure 3.8 shows an example for one scale.

- Most pixels are dark because filter responses are high only at finger positions. In a similar way for color histograms, we consider 100 positive bounding boxes for finding a standard deviation σ_f of pixel intensities for those responses to bounding boxes.

- We determine a histogram for every scale with 7 bins limited by 0, $0.25\sigma_f$, $0.50\sigma_f$, $0.75\sigma_f$, σ_f, $2\sigma_f$, and $3\sigma_f$. These values were determined experimentally. Frequency features for every bounding box are represented for 35 values coming from these 5 histograms. Figure 3.9 shows one example.

- Get a model of frequency histograms using a Kernel SVM for some labeled bounding boxes from training data. Figure 3.10 shows scores for positive and negative bounding boxes from training data. We can see that bounding boxes with a low color score (less than -0.95) and medium confidence (less than -0.65) can be dropped.

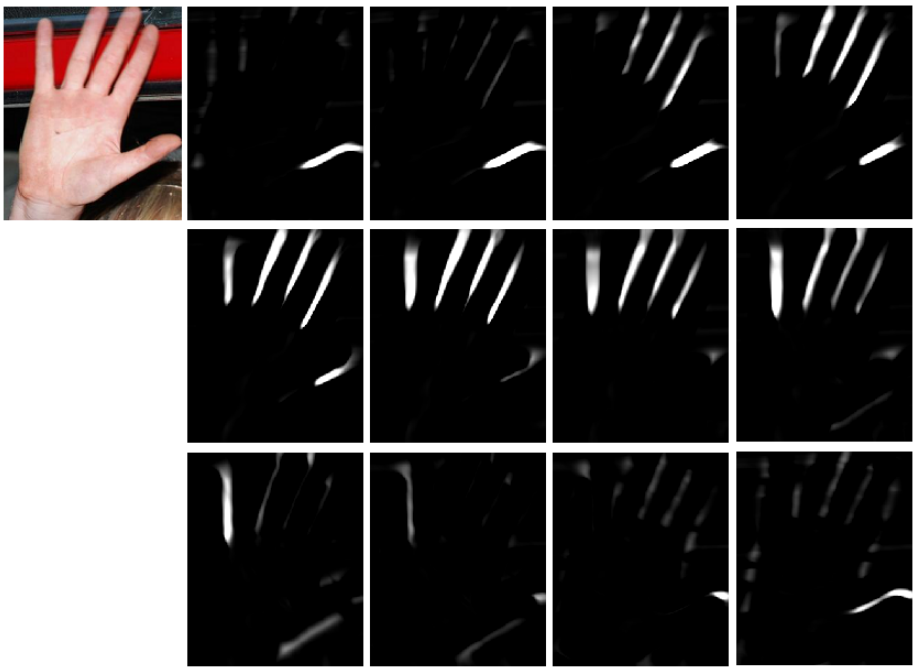

Figure 3.8: Directional filters responses for one scale where brighter color means high response in the respective angle.

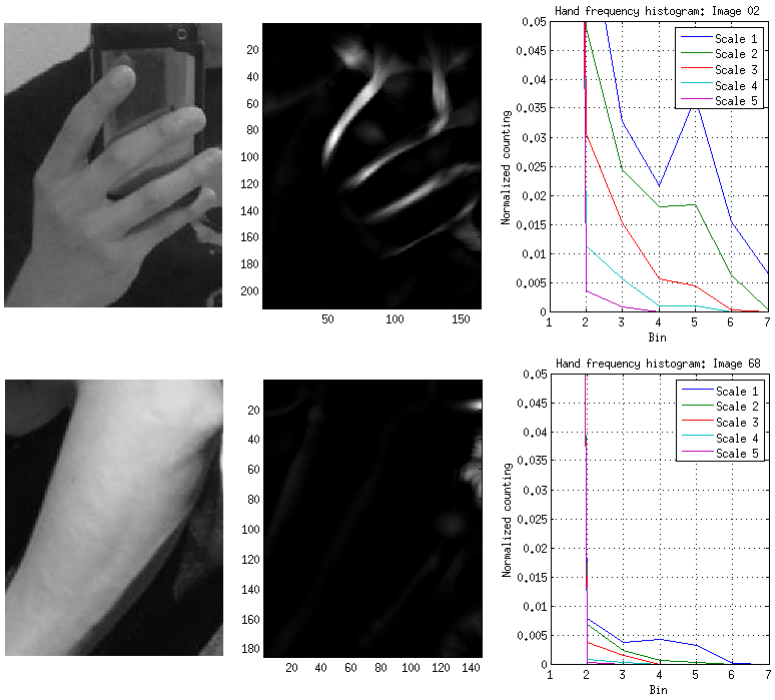

Figure 3.9: Histograms of energy for bounding boxes with hand (top) and no hand (bottom). Left: Grayscale bounding box. Center: image with maximum energy for one scale. Right: Energy histograms for every scale.

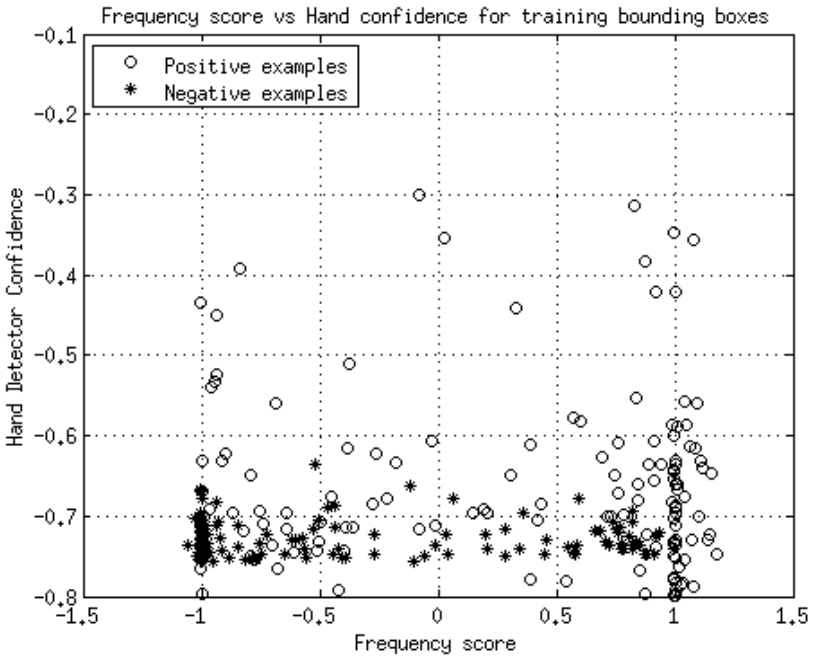

Figure 3.10: Positive and negative examples from training data

.

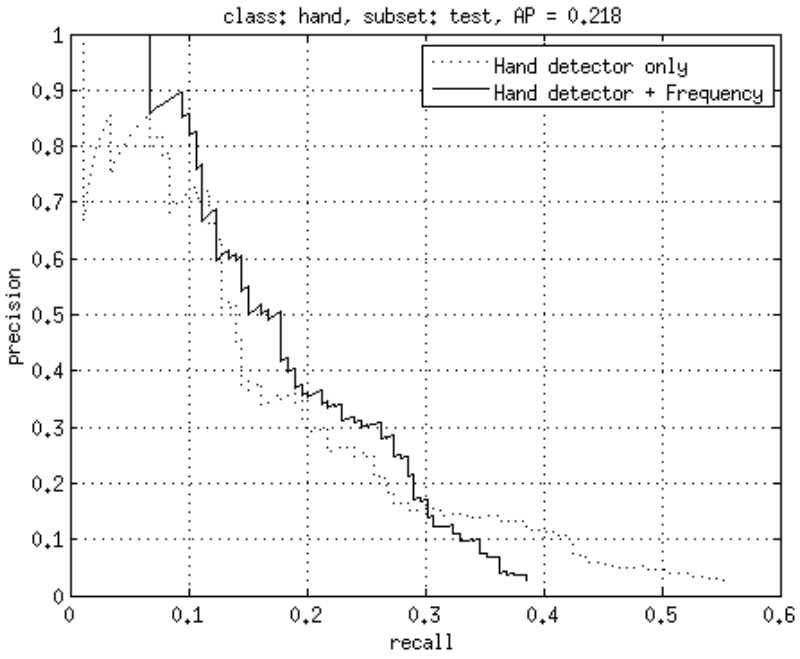

Figure 3.11: Precision-recall curve on test data including frequency features. The average precision for this detector is 0.218 while for detector 1 it is 0.215. Note the higher precision for recall less than 0.3

.

- Get a score for all bounding boxes from testing data using the model described in previous step and drop all bounding boxes with color score below -0.90.

The precision-recall curve after applying the frequency features to initial bounding boxes from testing data is shown in Figure 3.11. Notice that the average precision (0.218) is a little bit bigger than the original (0.215) and we have higher precision for low recall. The most important observation is that we have higher precision for recall from 0 to 0.3. This confirms that frequency features with directional filters helps to improve our precision for an important recall range.

3.4 Detector 4: Using Color and/or Frequency Features

This section describes a couple of detectors which uses both previous detectors based on color and frequency features. The first one considers the label given by detector 2 and detector 3, while the second one considers the label given by detector 2 or detector 3.

Both precision-recall curves are shown in Figures 3.12 and 3.13. The average precision for both are similar to that one given by detector1, however we also have a better precision for low recall.

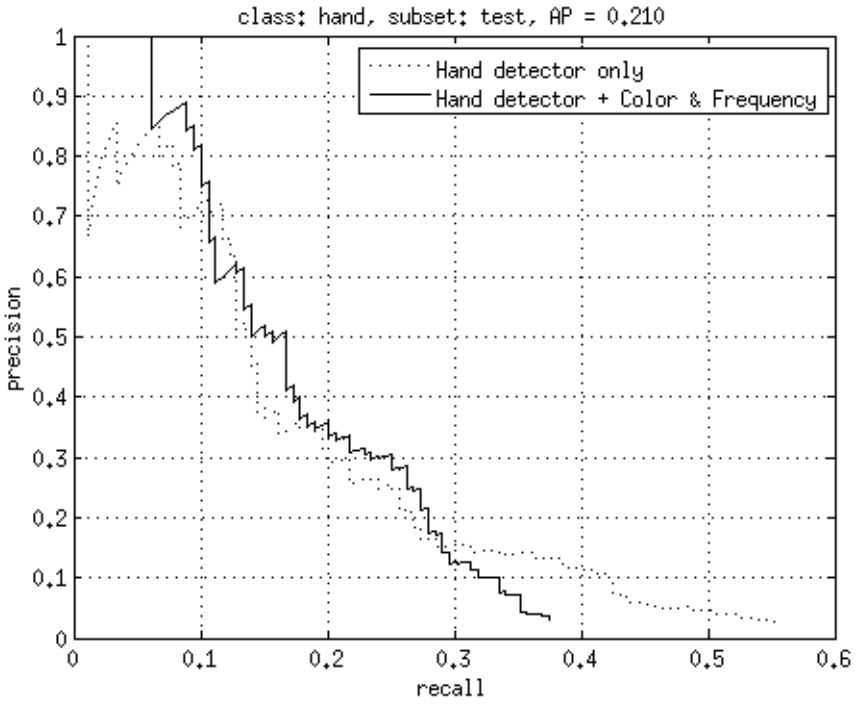

Figure 3.12: Precision-recall curve including color and frequency features. The average precision for this detector is 0.218 while the average precision for detector 1 is 0.215. Note the higher precision at low recall.

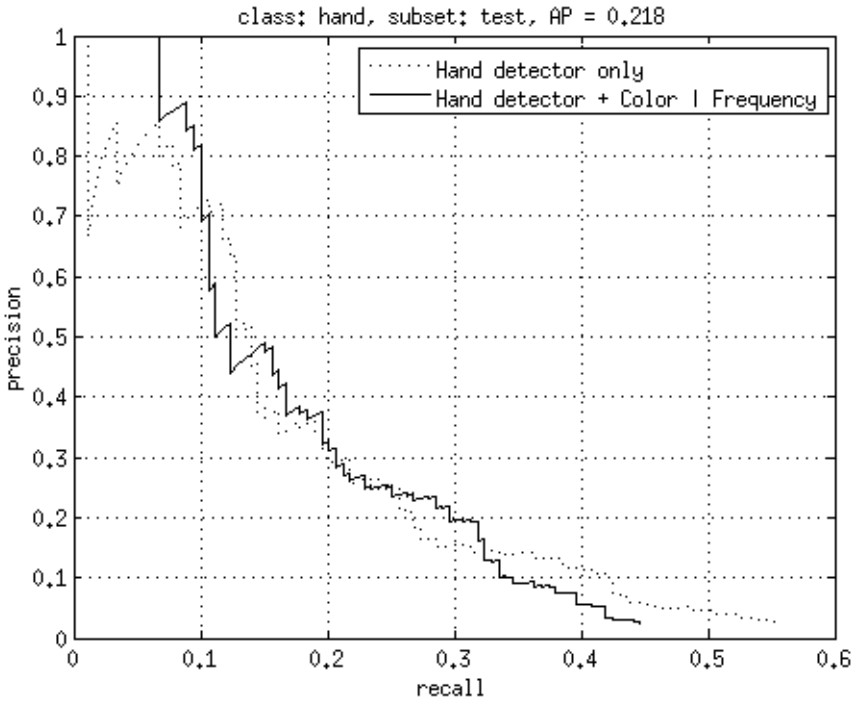

Figure 3.13: Precision-recall curve including color or frequency features. The average precision for this detector is 0.218 while the average precision for detector 1 is 0.215.

Chapter 4

Results

In this chapter we show a set of images and scores for bounding boxes from the testing data using our detectors. Table 4.1 shows the number of bounding boxes and the average precision for each detector we have designed. Initially we have 3693 bounding boxes given by Detector 1. Then, next detectors drop some of these bounding boxes based on the respective score and threshold as we explain in chapter 3.

Detector	# of bounding boxes	Average precision
Detector 1: baseline	3693	0.215
Detector 2: Det1 + color	3323	0.202
Detector 3: Det1 + frequency	2447	0.218
Detector 4a: Det1 + color & frequency	2258	0.210
Detector 4b: Det1 + color | frequency	3512	0.218

Table 4.1: Results for all hand detectors

We present 120 bounding boxes and their respective scores, in four groups of 30, sorted by hand detector (Detector 1) confidence:

- Set 1: Images and scores for bounding boxes 1 to 30. Figures 4.1 and 4.2.
- Set 2: Images and scores for bounding boxes 31 to 60. Figures 4.3 and 4.4.
- Set 3: Images and scores for bounding boxes 61 to 90. Figures 4.5 and 4.6.
- Set 4: Images and scores for bounding boxes 91 to 120. Figures 4.7 and 4.8.

All bounding boxes images were rescale in order to have the same height for visualization. Notice that color and frequency score has a relation with the information on their respective bounding box. For example, for bounding boxes 25, 26 and 28 (from Figure 4.1) with no hands, the color score is relative high (see Figure 4.2) because the central part of them correspond to skin color from body parts, however frequency score is low because the contrast change on those bounding boxes is less than that one in presence of fingers.

Figure 4.1: Bounding boxes 1 to 30 from testing data (Set 1), from left to right and top to bottom.

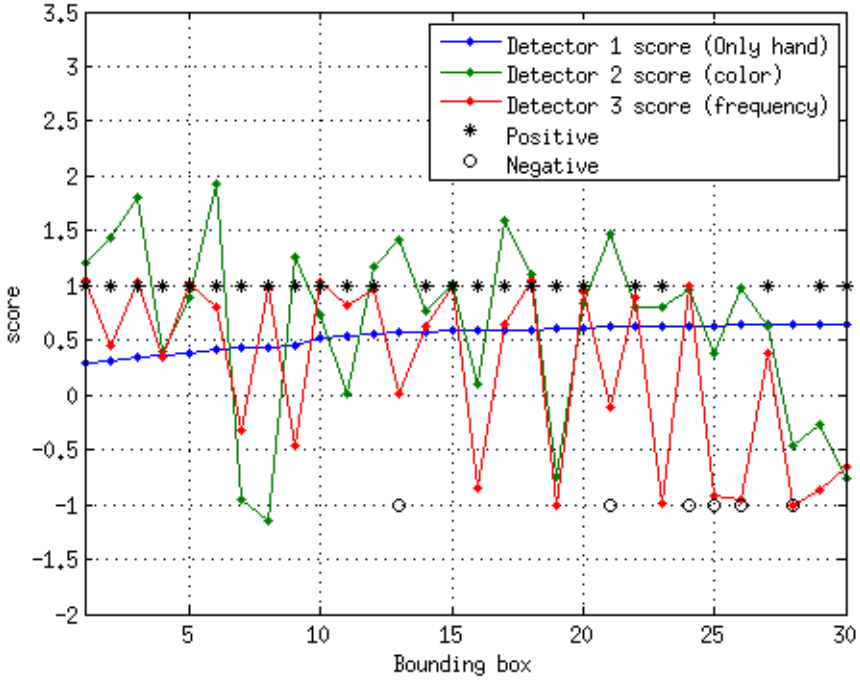

Figure 4.2: Scores for bounding boxes 1 to 30 show in Figure 4.1.

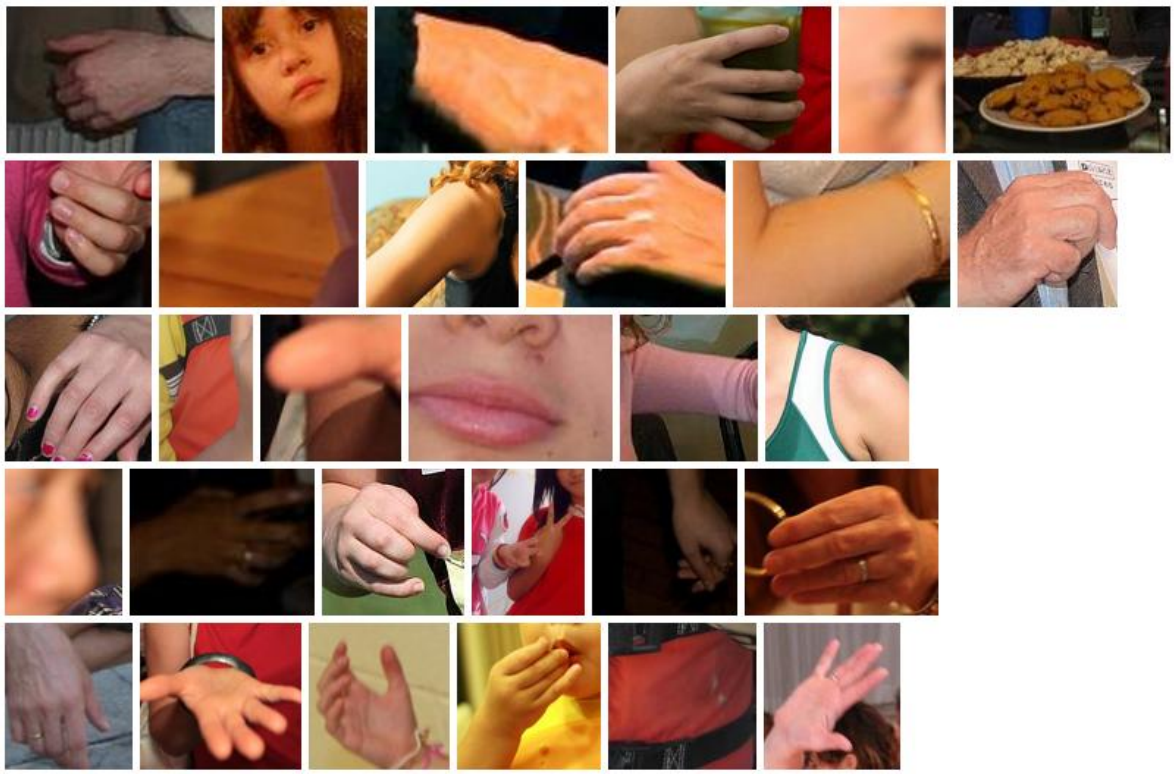

Figure 4.3: Bounding boxes 31 to 60 from testing data (Set 2), from left to right and top to bottom.

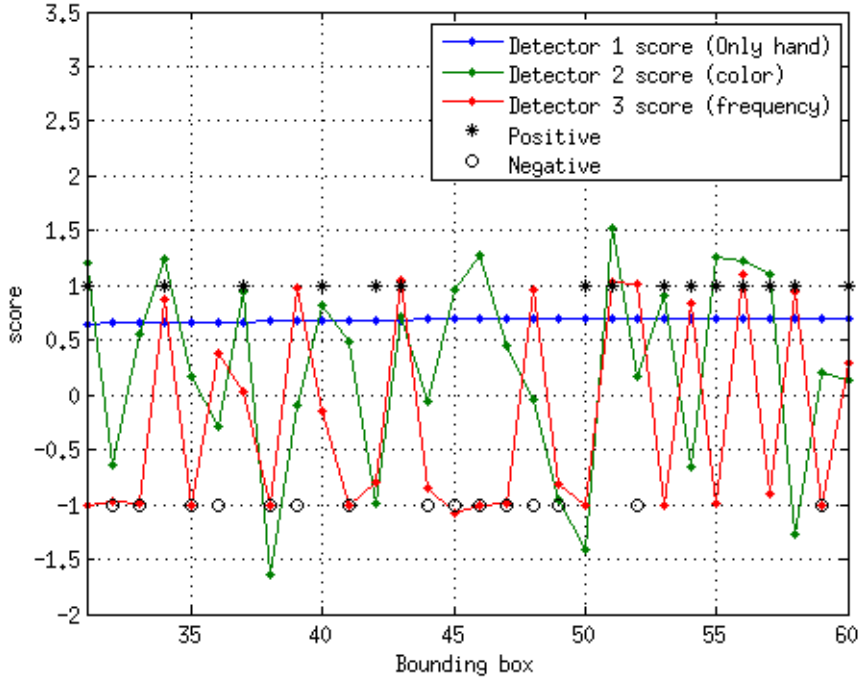

Figure 4.4: Scores for bounding boxes 31 to 60 show in Figure 4.3.

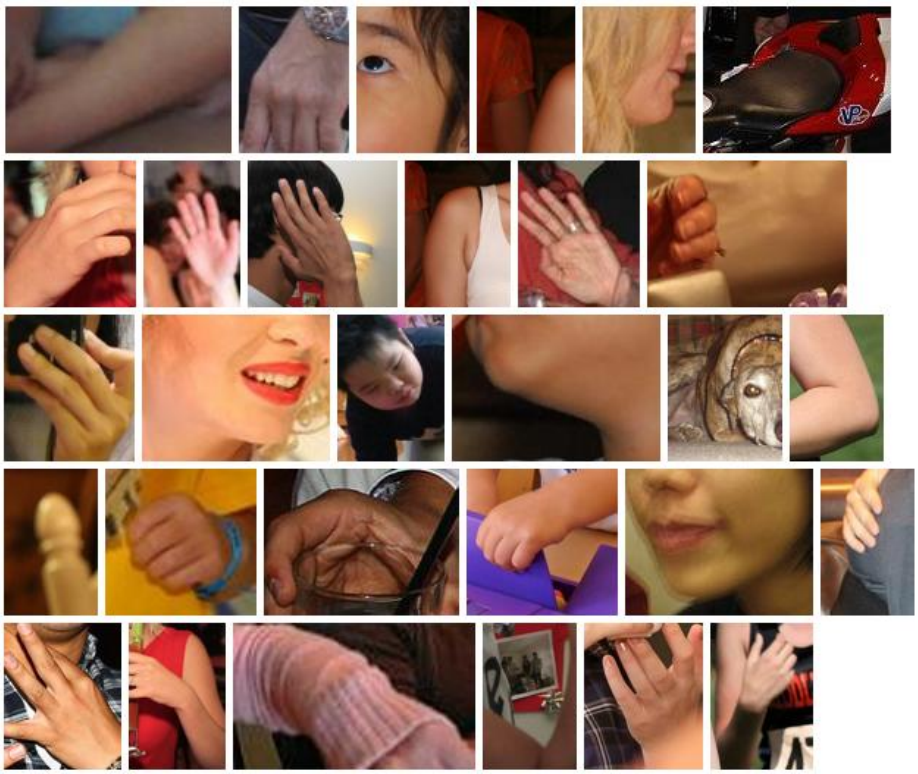

Figure 4.5: Bounding boxes 61 to 90 from testing data (Set 3), from left to right and top to bottom.

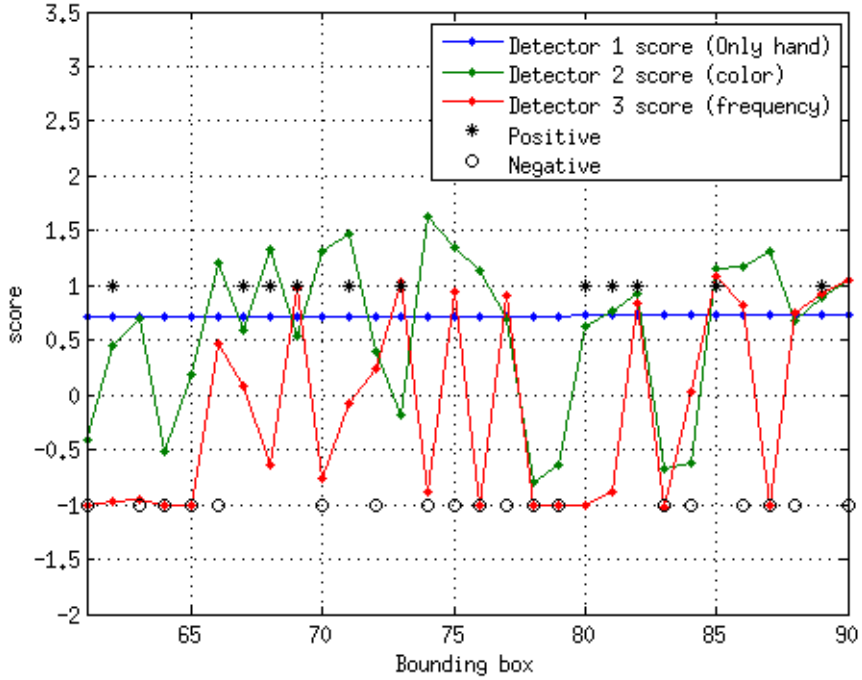

Figure 4.6: Scores for bounding boxes 61 to 90 show in Figure 4.5.

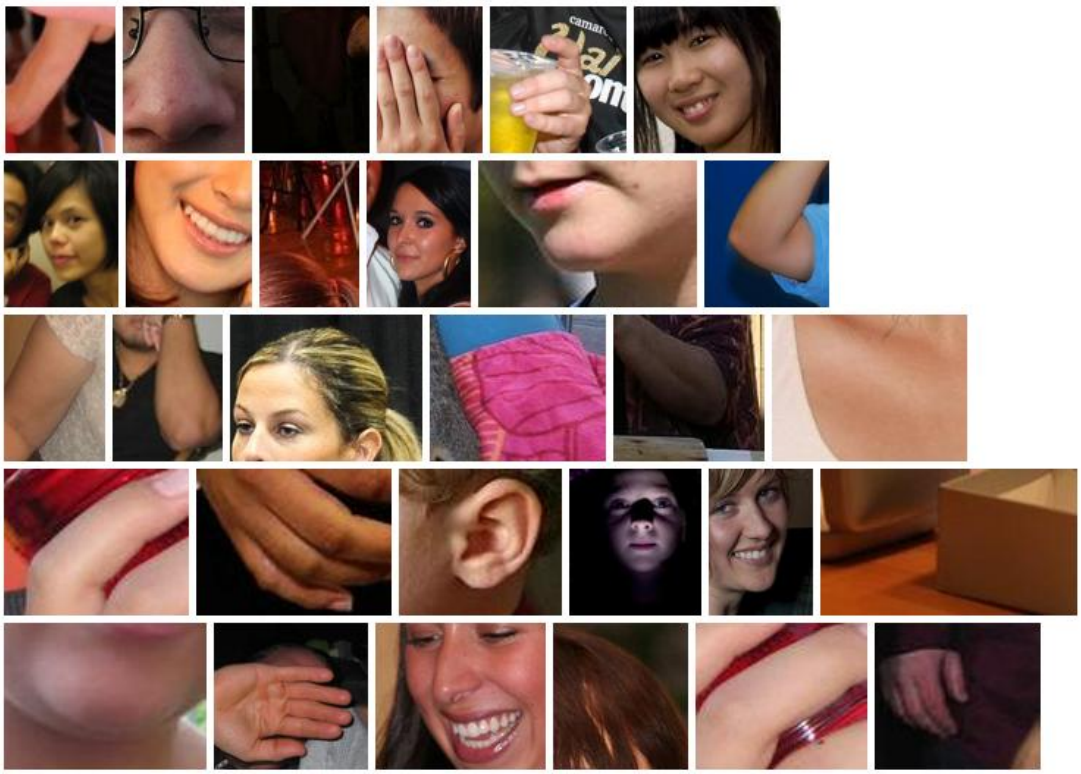

Figure 4.7: Bounding boxes 91 to 120 from testing data (Set 4), from left to right and top to bottom.

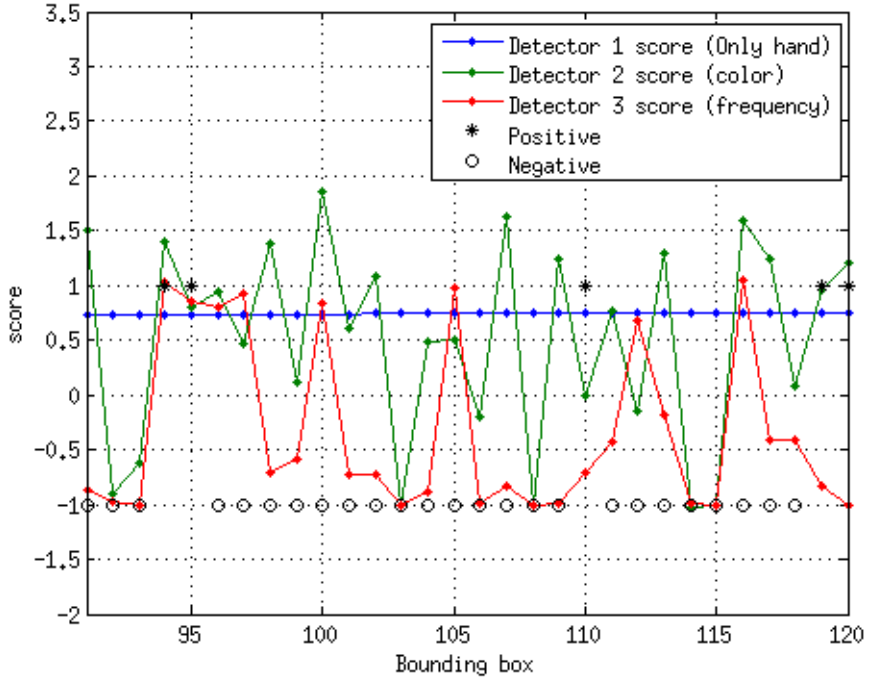

Figure 4.8: Scores for bounding boxes 91 to 120 show in Figure 4.7.

Chapter 5

Conclusions

The main conclusions from this work are:

- Using higher resolution images for the baseline hand detector gives us an average precision around 0.21, which is higher than state-of-the-art 0.104 reported by OXFORD SBD on Person Layout Results for VOC2010. Even when both are not directly comparable because we are working with different dataset, we state that resolution matters. Bigger images give us more details for learning and detecting hands.

- We have shown that both color and frequency features help by getting higher precision for low recall even when the average precision is still similar to the baseline. The baseline detector give us many false positives that we can drop using color or frequency criteria; however the average precision shows we also are dropping some true positives.

- Based on precision-recall curves from frequency and color features working alone with the baseline detector, we conclude that frequency features are more helpful than color ones.

- Kernel SVM got better result than linear SVM on modeling and evaluating colors or frequencies histograms from bounding boxes.

- Results suggest where may be more than one 'type' of hand image, boosting by down-weighting the hands that we detect may make it possible to detect other types accurately.

- In order to improve these results, future work should be focused on more training/testing data, feature and parameter adjusting for color and frequency histograms, and evaluation of new features.

References

[1] Vassilis Athitsos and Stan Sclaroff. Estimating 3d hand pose from a cluttered image, 2003.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[3] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/∼pff/latent-release4/.

[4] S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[5] Siddharth Swarup Rautaray and Anupam Agrawal. Article: A real time hand tracking system for interactive applications. *International Journal of Computer Applications*, 18(6):28–33, March 2011. Published by Foundation of Computer Science.

[6] James Rehg and Takeo Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In *In Proceedings of the workshop on Motion of Non-Rigid and Articulated Bodies*, pages 16–22. IEEE Computer Society Press, 1994.

[7] Ilkay Ulusoy and Christopher Bishop. Comparison of Generative and Discriminative Techniques for Object Detection and Classification. pages 173–195. 2006.

[8] Miaolong Yuan, Farzam Farbiz, Corey Mason Manders, and Ka Yin Tang. Robust hand tracking using a simple color classification technique. In *Proceedings of The 7th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '08, pages 6:1–6:5, New York, NY, USA, 2008. ACM.